VIDEO HANDOVER FOR RETRIEVAL IN A UBIQUITOUS ENVIRONMENT USING FLOOR SENSOR DATA

Gamhewage C. de Silva, T. Yamasaki, T. Ishikawa, Kiyoharu Aizawa
Department of Frontier Informatics, University of Tokyo
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan
{chamds, yamasaki, issy, aizawa}@hal.k.u-tokyo.ac.jp

ABSTRACT

A system for retrieving video captured in a ubiquitous environment is presented. Data from pressure-based floor sensors are obtained as a supplementary input together with video from multiple stationary cameras. Unsupervised data mining techniques are used to reduce noise present in floor sensor data. An algorithm based on agglomerative hierarchical clustering is used to segment footpaths of individual persons. Video handover is proposed and two methods are implemented to retrieve video and key frame sequences showing a person moving in the house. Users can query the system based on time and retrieve video or key frames using either of the handover techniques. We compare the results of retrieval using different techniques subjectively. We conclude with suggestions for improvements, and future directions.

1. INTRODUCTION

Video retrieval has been a fast growing research area in the recent few years. Despite the substantial amount of research, video retrieval is still a difficult task other than for highly structured video. Accessing and retrieving relevant video segments from unstructured video becomes especially important for electronic chronicles[1].

Video retrieval from ubiquitous environments poses additional challenges. Larger and more real-life environments with a large number of cameras are being built. The amount of video is large, and increasing with time. The content is much less structured compared to a single video from a specific category. Retrieval is required at multiple levels of granularity, not merely as a summary.

One difficult task in video retrieval from ubiquitous environments is to retrieve video that corresponds to a particular person, or event. Switching between videos from multiple cameras to show a particular person, we call *video handover*, is challenging.

Given the large amount of image data and the current state of the art of image processing algorithms, it is evident that video retrieval based solely on image data is a difficult task. Therefore it is desirable to make use of supplementary data from other sensors for easier retrieval.

This paper presents our work on video retrieval using video and sensory data from a ubiquitous environment. Unsupervised data mining algorithms have been used to reduce noise in data and retrieve video corresponding to people in the environment. The results are used to create a video chronicle that can be queried interactively.

2. RELATED WORK

A fair number of smart and ubiquitous environments have been built during the last decade. The *Ubiquitous Sensor Room* [2], *Aware Home* [3], and *CHIL* [4] are examples of more recent and ongoing projects.

Although there exists a fair amount of research on video retrieval, most of the work deals with specific content. Examples are sports video summarization [5], and analysis of news [6]. Most of the existing works use audio or text as a supplementary input for retrieval. Life log video captured by a wearable camera has been dealt with by using supplementary context information [7]. Context such as location, motion, time etc. is used for retrieval.

3. ENVIRONMENT AND SENSORS

The environment selected for this work is the *Ubiquitous Home* [8] at Keihanna Human Info-Communications Research Center, Kyoto, Japan. This is a two-bedroom house equipped with cameras and pressure-based floor sensors. We use the data from floor sensors for summarization and retrieval of video from the cameras. Figure 1 illustrates the layout and floor sensor arrangement of the ubiquitous home.

Ceiling-mounted stationary cameras record images at the rate of 5 frames/second. Point-based floor sensors are spaced by 180mm on a square grid. Their coordinates are specified in millimeters, starting from the bottom left corner of Figure 1. The pressure on the sensors is sampled at 6Hz. The state of a sensor, which is initially 0, changes to 1 when the pressure on a sensor crosses a specific threshold. Only the state transitions are recorded.

0-7803-9332-5/05/$20.00 ©2005 IEEE

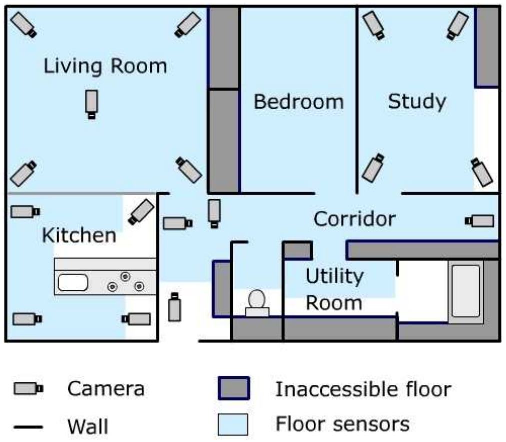

Figure 1: Layout and arrangement of sensors

A few problems arise from the installation and interfacing of floor sensors. A single footstep can activate 1 to 4 sensors. Damping due to flooring above the sensors can cause a delay in activation. Due to the low sampling rate, accuracy of the timestamps is low. The floor sensor data was found to contain two types of noise. One consists of pairs of state transitions with a time interval of 30-60 ms between them. These are caused by footsteps on nearby sensors. The other occurs when the sensors are loaded by a relatively small weight such as a leg of a stool. This noise is periodic, alternating between states 0 and 1.

For acquisition of experimental data, two voluntary subjects spent 3 days in the ubiquitous home. Data were acquired for approximately 9 hours each day. The subjects performed simple household tasks, and had meetings with up to 3 visitors. The actions of the subjects were not pre-planned, to ensure that the data are sufficiently general to be considered as those from a real-life situation. The images and the sensor data were stored with timestamps.

4. STEP SEGMENTATION

4.1 Preprocessing Data

Table 1 shows a subset of the recorded floor sensor data. The entries are ordered in time. The placing and removal of a foot on the floor will result in one or more pairs of lines. However the pairs may not be contiguous, as demonstrated by highlighted rows.

Table 1: Format of floor sensor data

Timestamp	X	Y	State
2004-09-03 09:41:20.64	1920	3250	1
2004-09-03 09:41:20.96	2100	3250	1
2004-09-03 09:41:20.96	1920	3250	0
2004-09-03 09:41:21.60	2100	3250	0

We use a pair-wise clustering algorithm to produce a single data entry, referred to as a *sensor activation*, for each pair of lines of input data. Table 2 shows sensor activations corresponding to Table 1. The timestamps are encoded in to a numeric format for ease of programming. The highlighted entry in Table 2 corresponds to the highlighted pair of rows in Table 1.

Table 2: Format of sensor activation data

Start time	End time	Duration	X	Y
34880.640	34880.968	0.328	1920	3250
34880.968	34881.609	0.641	2100	3250

Kohonen Self Organized Maps (SOM) were used for noise reduction. Both types of noise mentioned in Section 3 formed clusters in SOM's, enabling easy removal.

4.2 Hierarchical clustering

The next step is to divide the data into subsets corresponding to each person. This is performed using a technique based on Agglomerative Hierarchical Clustering (AHC). Figure 2 is a visualization of this process. The grid corresponds to the floor sensors. Activations that occurred later are indicated with a lighter shade of gray. Nearest neighbor clustering is performed in 3 levels with different distance functions defined as appropriate.

In the first level, sensor activations caused by a single footstep are combined. The distance function is based on connectedness and overlap of durations. For the second level, the distance function is based on the physiological constraints of walking, such as the range of distances between steps, the overlap of durations in two footsteps, and constraints on direction changes. However, due to the low resolution and the delay in sensor activations, the floor sensor data are not exactly in agreement with the actual constraints. Therefore, we obtained statistics from a data set corresponding to a single walking person and used the statistics to identify the ranges of values.

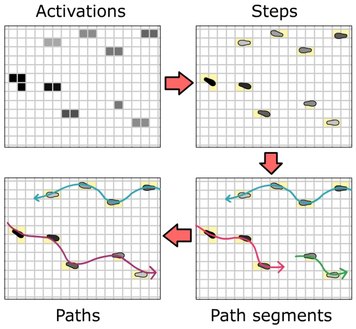

Figure 2: Step segmentation

The third step compensates for fragmentation of individual paths due to the absence of sensors at some areas, long steps etc. Context data such as the locations of the doors and furniture, and information about places where floor sensors are not installed, are used to for clustering.

In addition to paths, data regarding persons entering and leaving the house are extracted. These consist of timestamps and key frames from cameras near the entrance to the house.

5. VIDEO HANDOVER

Our intention is to automatically create a video clip showing a given person as he moves within the ubiquitous home. With more than one camera to view a given location, it is necessary to choose the most appropriate camera. There can be different, sometimes conflicting criteria. Examples are, to obtain a frontal view of the person in most of the sequence, and to have a smaller number of transitions between cameras.

In this work we implement 2 methods for video handover. In the first, we select the camera to view a person based only on his current position. In the second, we try to obtain a frontal view of the person where possible, by calculating the direction of his/her movement.

5.1 Camera View Model

A view model, as shown in Figure 3, was constructed for each camera. The projection of the optical axis of the camera on the XY plane, ***V*** is stored as a unit vector. The visibility of a human standing at the location of each floor sensor is represented by the value of 1. This mapping was created manually by observing images obtained during the experiment. The set of models can be looked up to identify cameras that can see a person at a given position.

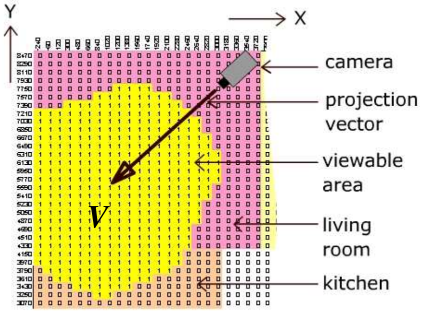

Figure 3: Camera mapping

5.2 Position-based handover

In this method, only the current position of a walking person is considered when selecting a camera to view that person. If the person can be seen from the previous camera, then that camera is selected. If not, the mapping for each camera is examined in a predetermined order. The first match is selected.

5.3 Direction-based handover

The direction vector of a walking person at step *P*, $\boldsymbol{D}_P$ is estimated by:

$$\boldsymbol{D}_P = \quad \boldsymbol{D}_{P\text{-}1} + (1 - \quad)(\boldsymbol{X}_P - \boldsymbol{X}_{P\text{-}1})$$

Here, $\boldsymbol{X}_P$ is the position vector of the step *P*. The value of   has been empirically set to 0.7 to obtain a relatively smooth direction with steps. The camera to be used is selected by evaluating the scalar product $\boldsymbol{V}.\boldsymbol{D}_P$ for each camera.

5.4 Retrieval of video and key frames

After determining the camera to be used at each step, it is straightforward to retrieve video from that camera, using the timestamps. In order to provide a summary of each person’s stay, a set of key frames is extracted from the video clip. A key frame is acquired every time the camera is changed and once every 5 seconds.

5.5 Querying for results

The results are stored in a database to be queried through a graphical user interface. A query is initiated by entering the time interval for which the summary is required. For the people who entered or left the house during the time interval, the key frames showing them entering or leaving the house will be displayed with timestamps. For those who entered the house before the specified time interval and remained inside, a key frame at the start of the time interval is displayed. By clicking each key frame, it is possible to retrieve a video clip or a sequence of key frames showing each person using either of the handover methods, according to user’s choice.

6. RESULTS

The results of step segmentation are generally accurate despite the presence of noise, delays, and low resolution. There are some cases of swapping of paths between two persons. Prediction based on the current position and direction is a possible refinement.

Video clips obtained using position-based handover have fewer transitions than those obtained using direction-based handover. For direction-based handover, the calculated gradient is not a robust measure of direction when a person sits and makes foot movements or takes a step back. We are investigating the possibility of using a weighted function of both direction and distance, for improved handover.

Figures 4 and 5 show key frame sequences corresponding to a small time interval, extracted using position-based handover and direction-based handover respectively. The person being tracked has been marked by rectangles. It is evident that the frame sequence for direction-based handover consists of more key frames, though not necessarily more informative.

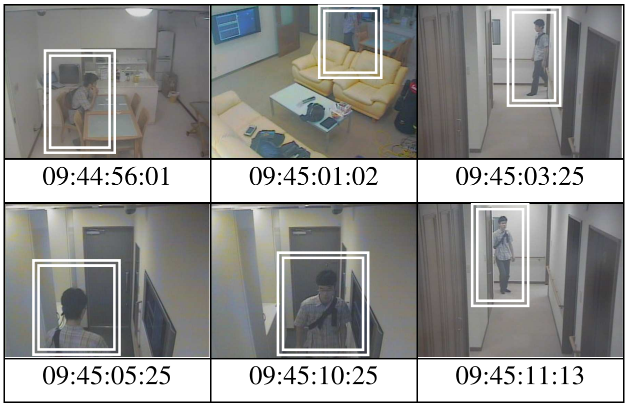

Figure 4: key frames by position-based handover

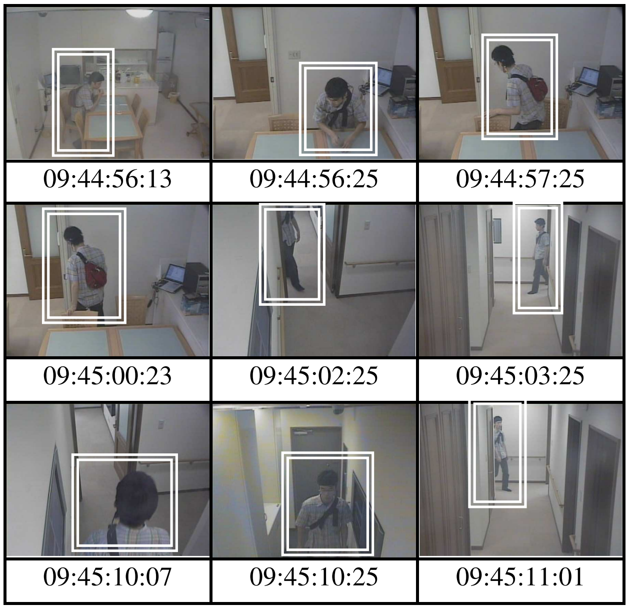

Figure 5: key frames by direction-based handover

7. CONCLUSION

A system for video retrieval for a ubiquitous environment has been presented. Data from pressure-based floor sensors are mined using Kohonen SOM's and hierarchical clustering to achieve accurate segmentation of paths taken by different persons. Two different video handover techniques have been implemented for retrieval of video and key frames corresponding to each person. The results can be accessed interactively with simple queries.

8. FUTURE WORK

We plan to extend the capability of the system to recognize and retrieve video corresponding to higher-level actions. A more detailed evaluation of step segmentation and video handover is in progress. Integration of retrieved results with a wearable life-log will enable viewpoints of the event by both the person himself and the environment.

9. ACKNOWLEDGMENTS

ByoungJun Oh participated in the experiment with the authors. This work was supported by the National Institute of Information and Communications Technology, Japan.

10. REFERENCES

[1] Ramesh Jain, “Multimedia Electronics Chronicles”, http://jain.faculty.gatech.edu/media_vision/eChronicles.pdf, Georgia Institute of Technology, USA.

[2] Department of Interaction Media, http://www.mis.atr.jp/~megumu/IM_Web/MisIM-E.html#usr, ATR MIS Laboratories, Kyoto, Japan.

[3] Abowd, G. A. Bobick, I. Essa, E. Mynatt, and W. Rogers, "The Aware Home: Developing Technologies for Successful Aging”, In proc. AAAI Conference 2002, Canada, July 2002. (2002).

[4] CHIL – Computers in the Human Interaction Loop, http://chil.server.de/servlet/is/101/, Interactive System Labs, Universität Karlsruhe (TH), Germany.

[5] Yong Rui; Anoop Gupta; Alex Acero; Automatically Extracting Highlights for TV Baseball Programs, Proc. ACM Multimedia, Oct. 2000, Los Angeles USA, Pages 105 –115

[6] Yuh-Lin Chang; Wenjun Zeng; Kamel, I. Alonso, R., Integrated image and speech analysis for content-based video indexing, Multimedia Computing and Systems, In Proc. of the Third IEEE International Conference on, 1996, Pp. 306 –313.

[7] T. Hori, K. Aizawa, Capturing Life Log and Retrieval based on Context, In Proc. IEEE ICME 2004, June 2004.

[8] Ubiquitous Home, http://www2.nict.go.jp/jt/a135/eng/research/ubiquitous_home.html, NICT, Japan.